基于混合式注意力机制的语音识别研究 *

李业良，张二华†，唐振民

(南京理工大学 计算机科学与工程学院，南京 210094)

摘 要：为了解决语音识别中基于卷积位置信息的混合式注意力机制无法提取长期有效位置信息的问题，提出了一种捕捉长期有效位置信息的新型混合式注意力机制。首先，对当前时刻生成的注意力得分作卷积来提取多通道特征图，并通过全局平均池化来得到恒定维度的特征向量；接着，引入长短期记忆网络(long short-term memory，LSTM)单元作为外部记忆模块，并以生成的特征向量作为输入，生成下一时刻的位置信息向量；最后，结合经典的 LAS(listen，attend and spell)模型来验证提出方案的有效性。实验结果表明，提出方案能充分考虑过去多个时刻的注意力得分。相对于基于卷积位置信息的 LAS 模型，提出方案在纯净和含噪语音数据集上取得的标签错误率分别减少了 1.8%和 2.21%。

关键词：卷积；注意力机制；全局平均池化；LSTM；LAS 模型

中图分类号：TP391.4 doi: 10.19734/j.issn.1001-3695.2018.06.0492

Research on speech recognition based on hybrid attention mechanism

Li Yeliang, Zhang Erhua†, Tang Zhenmin

(*School of Computer Science & Engineering*, *Nanjing University of Science & Technology*, *Nanjing* 210094, *China*)

Abstract: In speech recognition, the convolution location-based hybrid attention mechanism can not extract location information that can be valid over long term. This paper proposed a new hybrid attention mechanism to solve this problem. Firstly it convolved with the attention score generated for the current time to extract multi-channel features, followed by obtaining the feature vectors of constant dimensions via global average pooling. Then it introduced a LSTM (Long Short-Term Memory) unit as the external memory module and used the generated feature vectors as input to generate the location vectors for the next time point. Finally this paper used the classic LAS (Listen, Attend and Spell) model to verify the effect of the new hybrid attention mechanism. Experiment results show that the new hybrid attention mechanism can take full consideration of the attention scores at many past time points. Compared to the convolution location-based LAS model, the label error rate of the proposed scheme on pure and noisy speech datasets is reduced by 1.8% and 2.21%, respectively.

Key words: convolution; attention mechanism; global average pooling; LSTM; LAS model

0 引言

近年来，深度神经网络(deep neural network，DNN)在图像、机器翻译、语音识别领域取得了广泛的应用[1]。过去的深度神经网络，作为一个组成单元通常与隐马尔可夫模型(hidden Markov model，HMM)结合组成 DNN-HMM 声学模型[2~4]，但这些组成单元却是分别独立训练的，没有考虑单元之间的相互关系来进一步提高模型性能。最近，端到端的语音识别模型由于其架构简单、训练方便而受到了广泛关注。不同于 DNN-HMM 模型，其不需要输入数据和给定标签在时间上一一对齐，并且其本身作为一个整体架构来进行训练优化，故使得语音识别率得到很大的提升。目前主流端到端语音识别模型有两种：一种是 CTC(connectionist temporal classification)[5]模型；另外一种是基于注意力机制[6]的编码器—解码器模型。基于注意力机制的编码器—解码器模型是目前语音识别领域的一个研究方向。一般来说，以 LSTM[7]或 GRU(gated recurrent unit)[8]这类递归神经网络作为编码器和解码器，用编码器来处理变长的特征序列生成隐含状态序列；在解码阶段，通过引入注意力机制，解码器每一时刻的输出直接利用编码器每一时刻的隐含状态信息，最终生成相应的标签。

文献[9]给出了基于注意力机制的编码器—解码器模型在语音识别领域的第一个结果，它把逐帧提取的 fMLLR(feature space maximum likelihood linear regression)特征序列作为编码器的输入，音素序列作为输出，最终在 TIMIT 数

收稿日期：2018-06-03；修回日期：2018-07-25 基金项目：军委装备发展部“十三五”装备预研基金资助项目

作者简介：李业良（1993-），男，广东佛山人，硕士研究生，主要研究方向为深度学习、语音识别；张二华（1967-），男（通信作者），副教授，博士，主要研究方向为语音信号处理、语音分离（zherhua@163.com）；唐振民（1961-），男，教授，博士，主要研究方向为智能机器人系统技术、图像处理与目标识别.

据库取得 18.57%的音素错误率。文献[10]给出了 LAS 框架，其采用具有启发式的金字塔 BLSTM[7,11,12]结构作为编码器来对相邻时间帧的特征进行合并降维，减少注意力机制的计算量，加速网络的训练；同时在训练解码上，采用抽样前时刻的概率分布来取代传统的用前时刻真实标签作为下一时刻解码器输入[13]，这在一定程度上使得模型在推断中面对某时刻的错误输出能更加鲁棒地进行调整而不至于过度影响下一时刻的推断。

文献[14]提出了一种适用于语音识别任务的混合式注意力机制。与 LAS 模型中基于内容的注意力机制不同，通过引入对前一时刻的注意力得分的卷积结果作为该时刻的额外信息来调整注意力得分，使得模型对上下文信息的整合加进了对位置感知信息的考虑，最终取得性能上的提升。然而卷积层并没有提取长期依赖信息的能力，仅仅利用上一层提供的位置信息来对该时刻的注意力得分进行调整的做法仍有一定的偏差。

本文给出了基于卷积提取位置信息的注意力机制的一个改进方案，对注意力得分进行卷积和全局平均池化[15]生成固定维度的特征向量，并引进一个 LSTM 单元来对其进行处理从而生成下一时刻的位置信息。LSTM 单元的引入使得模型能充分考虑过去多个时刻的注意力得分，使得提供给下一时刻的位置信息更加合理和准确。

1　基于注意力机制语音识别模型

本章将介绍基于注意力机制的语音识别模型的基本架构及经典的混合式注意力机制。

1.1　LAS 模型

LAS 是一个实现音频特征序列到字母序列转录的基于注意力机制的基本模型，其通过识别出的字母序列的拼接得到对应的文本。具体来说是从音频数据中提取梅尔滤波器特征序列 $\boldsymbol{x}=(x_1,x_2,\ldots,x_T)$作为模型输入，以字符序列 $\boldsymbol{y}=(sos,y_1,\ldots,y_S,eos)$作为模型的输出，其中 $y_i\in\{a,b,\ldots,z,0,\ldots,9,period,space,$ $comma,apostrophe,unk\}$。在这里 sos 和 eos 分别作为句子起始和结束的记号，$space$ 和 unk 则分别指代空格和未知字符。模型训练的目标函数为使得输出序列关于输入序列的条件概率最大，按照概率链式法则公式表达如下：

$$P(\boldsymbol{y}\mid\boldsymbol{x})=\prod_i P(y_i\mid x,y_{i-1},\ldots,y_1) \tag{1}$$

LAS 模型由编码器和解码器两部分组成，文中分别称做 Listener 和 Speller。如图 1 所示，通常以金字塔结构堆叠 BLSTM 组成 Listener，从输入序列 $\boldsymbol{x}$ 生成隐含状态序列 $\boldsymbol{h}$：

$$\boldsymbol{h}=\text{Listen}(\boldsymbol{x}) \tag{2}$$

$$h_i^j=\text{pBLSTM}([h_{2i}^{j-1},h_{2i+1}^{j-1}],h_{i-1}^j) \tag{3}$$

其中：h_i^j 指的是第 j 层第 i 个时刻的 pBLSTM 单元输出的隐含状态，其中把第 j-1 层相邻时刻的隐含状态拼接起来的向量和前一个时刻的 pBLSTM 单元的输出状态作为当前单元的输入便生成当前的输出状态，如式(3)所示。式(2)给出了 Listener 的最终输出，即 pBLSTM 最后隐层输出的状态序列 $\boldsymbol{h}$。

Speller 是一个基于注意力的 LSTM Transducer[6,14]。Transducer 于每一步输出一个当前字符关于序列 $\boldsymbol{x}$ 及在这之前所有字符的条件分布，即 $P(y_i|\boldsymbol{x},y_{i-1},\ldots,y_1)$。该分布是一个关于解码器当前时刻输出状态 s_i 及上下文 c_i 的函数。解码器状态通过 LSTM 生成，而上下文 c_i 则由注意力机制产生，具体公式表达如下：

$$P(\boldsymbol{y}\mid\boldsymbol{x})=\text{AttendAndSpell}(\boldsymbol{y},\boldsymbol{h}) \tag{4}$$

$$s_i=\text{LSTM}([y_{i-1},c_{i-1}],s_{i-1}) \tag{5}$$

$$c_i=\text{AttentionContext}(s_i,\boldsymbol{h}) \tag{6}$$

$$P(y_i\mid x,y_{i-1},\ldots,y_1)=\text{CharacterDistribution}([s_i,c_i]) \tag{7}$$

其中：CharacterDistribution 是带有 MLP 的 softmax 回归，其以 s_i 和 c_i 的拼接向量作为输入。

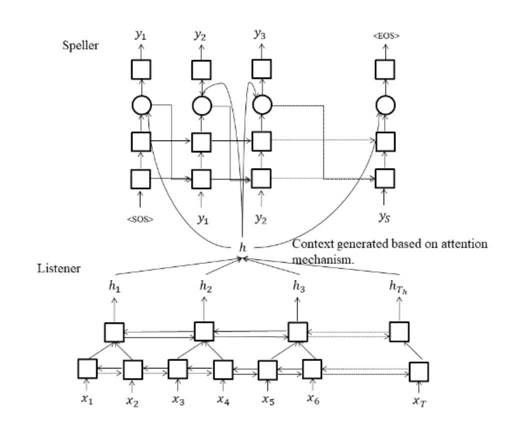


图 1　LAS 模型框架

Fig.1　LAS model framework

1.2　注意力机制

注意力机制使得解码器每一时刻预测输出用到的上下文信息是与当前输出有关系的上下文，即通过对编码器每一时刻的输出状态打分，然后根据得分对输出状态加权求和得到上下文。文献[14]给出注意力机制的一般表达形式：

$$\alpha_i=\text{Attend}(s_{i-1},\alpha_{i-1},\boldsymbol{h}) \tag{8}$$

$$c_i=\sum_{j=1}^{T_h}\alpha_{ij}h_j \tag{9}$$

由式(8)确定的注意力模型通常称为混合式注意力模型。将式(8)中的 α_{i-1} 忽略，就得到基于内容的注意力模型[16]。式(6)也是基于内容的注意力模型，但与式(8)不同，α_i 是由本时刻的输出状态 s_i 和编码器输出状态序列 $\boldsymbol{h}$ 决定的，这种叫 Luong 注意力[17]。基于内容的注意力模型存在一个缺陷，由于一段音频中存在大量的重复帧，所以生成的注意力得分会因为内容相近而值相近，可能会导致某一时刻的解码输出过于依赖相同含义的帧。为了解决这个问题，文献[14]引入了前一时刻的注意力得分的卷积结果作为位置信息来对该时刻的注意力得分进行调整。具体方法是增加一个矩阵 $\boldsymbol{F}$，并让它和前一时刻的得分做卷积：

$$f_i=\boldsymbol{F}*\alpha_{i-1} \tag{10}$$

然后再重新计算 α_i：

$$e_{ij} = w^T \tanh(\boldsymbol{W}s_{i\text{-}1} + \boldsymbol{V}h_j + \boldsymbol{U}f_{ij} + b) \tag{11}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_h} \exp(e_{ik})} \tag{12}$$

2 新型混合式注意力机制

上述提到的基于卷积提取位置信息的混合式注意力机制使得模型性能得到一定程度的提高，但这样做的局限性是卷积不能综合过去多个时刻的注意力得分来得到更准确的位置信息。与文献[14]不同，本文通过引入 LSTM 来解决这一问题。LSTM 是 RNN 的一种特殊类型，可以更好地学习长期依赖信息。LSTM 通过特意的设计来避免梯度消失的问题，记忆长期信息是 LSTM 的默认行为，而非需要付出很大的代价才能获得的能力。图 2 给出了 LSTM 单元的构造图。原生的 RNN 随着时间的推移，后面的时间节点对前面时间节点的感知力下降，这会导致长期历史信息无法充分利用。而 LSTM 通过引入“细胞”来更好地记住长期依赖的信息，同时引入“门”机制来对历史信息和当前输入信息进行筛选而产生更好的输出状态。文献[7]给出了常规的 LSTM，具体公式如下：

$$i_t = \sigma(\boldsymbol{W}_{xi}x_t + \boldsymbol{W}_{hi}h_{t\text{-}1} + b_i) \tag{13}$$

$$f_t = \sigma(\boldsymbol{W}_{xf}x_t + \boldsymbol{W}_{hf}h_{t\text{-}1} + b_f) \tag{14}$$

$$c_t = f_t c_{t\text{-}1} + i_t \tanh(\boldsymbol{W}_{xc}x_t + \boldsymbol{W}_{hc}h_{t\text{-}1} + b_c) \tag{15}$$

$$o_t = \sigma(\boldsymbol{W}_{xo}x_t + \boldsymbol{W}_{ho}h_{t\text{-}1} + b_o) \tag{16}$$

$$h_t = o_t \tanh(c_t) \tag{17}$$

其中：σ 指代 sigmoid 函数；i_t、f_t、o_t 分别指代输入门、遗忘门和输出门；c_t 和 h_t 别称为细胞状态和隐含状态，其中 h_t 作为 LSTM 在每一时刻的输出。

新型的混合式注意力机制具体公式表达如下：

$$e_{ij} = q^T \tanh(\boldsymbol{W}_s s_i + \boldsymbol{W}_h h_j + \boldsymbol{W}_m m_i + b) \tag{18}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_h} \exp(e_{ik})} \tag{19}$$

$$c_i = \sum_{j=1}^{T_h} \alpha_{ij} h_j \tag{20}$$

$$f_i = \boldsymbol{F} * \alpha_i \tag{21}$$

$$p_i = \text{GlobalAveragePooling}(f_i) \tag{22}$$

$$m_{i+1} = \text{LSTMUnits}(p_i, m_i) \tag{23}$$

其中：e_{ij} 计算的是解码器输出 s_i 和编码器某一时刻的输出状态 h_j 的相似度，与式(11)类似，本文还考虑了前一时刻提供的位置信息 m_i。通过对 e_{ij} 进行 softmax 正则化便可得到关于编码器第 j 个时刻输出的注意力得分 α_{ij}，最后利用该得分对编码器输出序列进行加权平均便可得到上下文 c_i。式(21)~(23)是本方案的创新点。由于编码器的输出状态序列长度是可变的，为了能从中提取有用的位置信息，本文采用卷积处理。多个大小为 5 的卷积核进行卷积操作 $\boldsymbol{F} * \alpha_i$ 提取多通道特征，再通过全局平均池化便可提取出固定维度的向量 p_i，最后使用 LSTMUnits 函数(式(13)~(17))便可生成下一时刻位置感知信息 m_{i+1}。

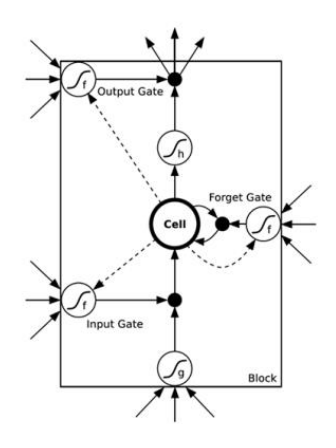


图 2 LSTM 单元构造图

Fig.2 Structure diagram of LSTM unit

3 解码和重新编码

与文献[10]不同的是，在推断阶段，本文仅考虑声学模型，即在给定输入音频的基础上寻找概率最高的字符序列：

$$\hat{y} = \text{argmax}_y \ln P(\boldsymbol{y} \mid \boldsymbol{x}) \tag{24}$$

传统的广度优先搜索策略能找到最优解码，但搜索空间非常大，容易造成内存溢出，本文采用 beam search[18] 算法进行解码。用记号 *sos* 作为解码器开始时刻的输入，并维持 β 大小的局部假设路径集合 beam。在每一个时刻，用词汇表中的每个单词扩大 beam 中的每条局部假设路径，然后根据目前的局部假设路径对其剪枝并保留概率最高的 β 条路径，直到某条路径以 *eos* 结尾，把它从 beam 中移出并加入完备假设集。

4 实验及分析

4.1 实验方法

本文分别在纯净和含噪的数据集进行两组对比实验来评估提出的方案。纯净语音数据来自于 Voxforge[19]，抽取同一人的 1 138 组语音样本作为数据集，并随机划分 910 组语音样本为训练集，剩余样本为测试集。在每一段语音数据中加入来自 NOISEX-92[20]数据库的工厂噪声便能生成相应的含噪语音，其中纯净语音信号和噪声信号的平均信噪比约为 7 dB。本实验的对比模型分别是 BLSTM-CTC 模型[11]、基于卷积位置信息的编码器—解码器模型[14]、LAS 模型和基于卷积位置信息的 LAS 模型。其中基于卷积位置信息的 LAS 模型是 LAS 模型的一个改进，即在注意力机制提取上下文信息的过程中加入上一时刻提取的卷积位置信息，加强该时刻提取的上下文的合理性。该模型是本实验比较的重点。本实验以每 25 ms 作为一帧（10 ms 帧移）提取 40 维的对数梅尔滤波器特征作为模型 Listener 的输入，并且对特征进行 z-score 标准化加快模型训练速度。

文本序列中所有的英文字母均转换为小写，解码部分考虑的符号为英文字母、数字、句号、逗号、空格、引号，其他符号用 *unk* 来表示。每段文本的开头和结尾分别添加记号 *sos* 和 *eos*。由于字母是一个离散值，所以采用独热编码将其映射到连

续空间而作为 Speller 的输入。

对于 Listener，使用节点数为 128 的 BLSTM 从输入特征序列中提取隐含状态序列，再用节点数为 256 的 pBLSTM 对其进行时间维度上的降维。对于 Speller，采用 2 层节点数为 128 的 LSTM 进行解码。模型参数以均匀分布 $u(-1.0,1.0)$进行初始化。为了避免深度神经网络中出现 internal covariate shift[21]，加速模型的收敛，对网络中每一层输入进行 layer normalization[22]变换。

本文采用 Adam[23]优化算法来进行模型的训练，其中超参数 β_1 和 β_2 均设为 0.9，学习速率则设为 0.003。采用指数衰减法来调节学习速率以使模型在训练后期更加稳定，其中衰减系数设为 0.55，衰减步数为 2 000。

在推断阶段，采用预测字母而不是真实字母作为下一时刻的输入。当输入的是一个错误的预测时，则可能会导致后续每个时刻都作出错误的预测，因为这可能是训练阶段没见过的状态分布。为了减缓这种影响，采用 Scheduled Sampling[13] 的训练方法，即在训练阶段以一定的采样概率使上一时刻的预测字母作为下一时刻的输入，在本实验中采样概率为 25%。为了使解码更加稳定，在测试集上采用 beam search 解码。

4.2　各模型的标签错误率对比

图 3 给出了不同集束宽度下的模型解码对比。可以看出，无论是纯净测试集还是噪声测试集，新的模型在每个 beam width 上都取得了最低的标签错误率(label error rate , LER) [5]。表 1 给出了所有实验的汇总结果。在纯净测试集上，新的模型取得了 22.06%的标签错误率，相比于基于卷积位置信息的的 LAS 模型减少了 1.8%；在噪声测试集上，新的模型取得了 24.97%的标签错误率，相比基于卷积位置信息的 LAS 模型减少了 2.21%。噪声的加入使得相近发音更加难以区分，基于卷积位置信息的 LAS 模型对上下文的提炼仅能依赖于隐含状态内容及前一时刻注意力得分提供的位置信息，故在噪声测试集上的标签错误率急速上升。通过使用 LSTM 单元替换卷积来提炼位置信息，充分发挥了 LSTM 长期记忆的优势，提供的位置信息更加合理和准确。新的模型变得更加鲁棒，每一步的解码，不过分关注内容相近的隐含状态，故仍能维持较低的标签错误率。

相对于基于卷积位置信息的编码器—解码器模型，LAS 模型在纯净测试集上取得的标签错误率减少了 1.5%；而在噪声测试集上，则高出 0.5%。这表明金字塔 BLSTM 的编码器结构通过融合相邻帧的输出状态能够减少解码过程中需要注意的特征信息，在干净的语音训练数据上能取得更好的性能提升。但该结构鲁棒性相对较弱，由于在注意力机制提取上下文信息的过程中缺少位置信息，使得模型把多余的状态信息考虑进来。而噪声的存在使得这些多余的状态信息对上下文信息影响较大，故在噪声测试集上性能大幅度下降。因此，理想的情况是把这样的编码器结构和混合式注意力机制结合起来提高模型最终性能，正如新的模型和基于卷积位置信息的 LAS 模型所展示的那样。

与原始的 LAS 模型和基于卷积位置信息的编码器—解码器模型相比，BLSTM-CTC 在纯净和噪声测试集上均取得更低标签错误率。与 CTC 框架不同，编码器—解码器框架每一步的解码不仅基于从语音信号中提取的上下文信息，还基于前一时刻的解码信息。若前一时刻的解码出现错误，将会影响该时刻的标签预测。编码器—解码器模型尽管采用 Scheduled Sampling 的训练方式减缓了这样的影响，但却无法完全消除。特别是在含噪测试集上，标签错误率上的差距被进一步拉大，这从侧面反映了在注意力机制中加入位置信息对正确解码的必要性。

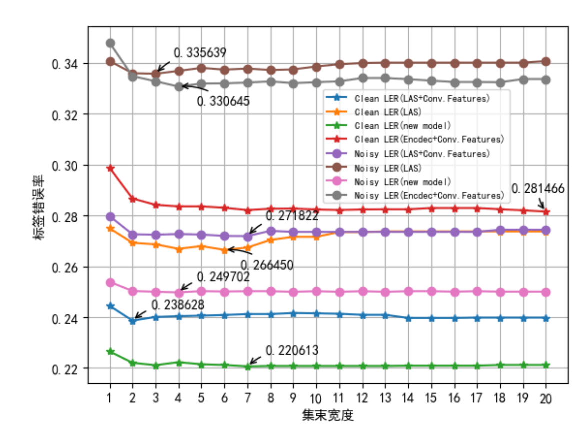


图 3　使用 beam search 在纯净和含噪测试集上获得的模型解码对比

Fig.3　Model decoding comparison using beam search on pure and noisy test sets

表 1　在纯净和含噪测试集上的标签错误率

Table 1　Label error rate on pure and noisy test sets

model	Clean LER	Noisy LER
BLSTM-CTC	24.14%	28.56%
Encdec+Conv.Features	28.15%	33.06%
LAS	26.65%	33.56%
LAS+Conv.Features	23.86%	27.18%
new model	**22.06%**	**24.97%**

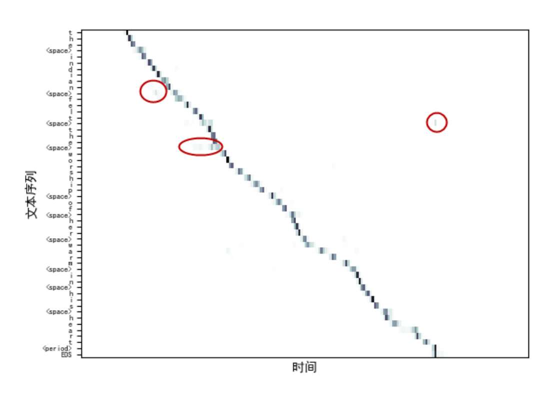
4.3　可视化注意力得分

图 4 分别给出了 LAS 模型,基于卷积位置信息的 LAS 模型和新的模型基于同一语音信号(图 4(d))的注意力得分可视化对比。不难发现，对于字符的解码，LAS 模型的注意力得分分布最稀疏(图(a)中的红圈所示)，这说明 LAS 模型的解码需要更多位置的内容信息的帮助，然而却没有考虑到位置的对齐，使得解码有点模棱两可。相对于 LAS 模型，其余两个模型的字符注意力得分分布则要更加集中，整体呈现出单调性，尤其是新的模型。这说明模型解码时不仅能找到相关内容信息，同时考虑了位置信息，这与前述的理论分析是一致的。

5　结束语

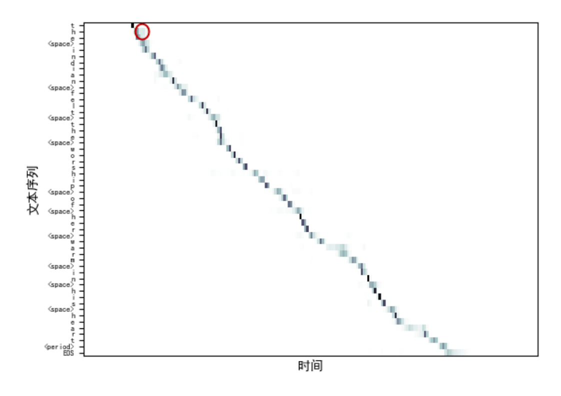
本文给出了基于卷积位置信息的混合式注意力机制的一个改进方案。具体做法是对当前时刻生成的注意力得分作卷积提取多通道的特征图（通道数是固定的），并再作全局平均池化来得到恒定维度的特征向量。引入一 LSTM 单元作为外部记忆模块，以生成的特征向量作为输入便能生成下一时刻的位置信息向量。本文结合经典的 LAS 模型来对新方法进行评估。实验

结果表明，新的模型在纯净和含噪的语音测试集上均取得最低的标签错误率，充分反映了 LSTM 对长期位置信息的记忆能力。



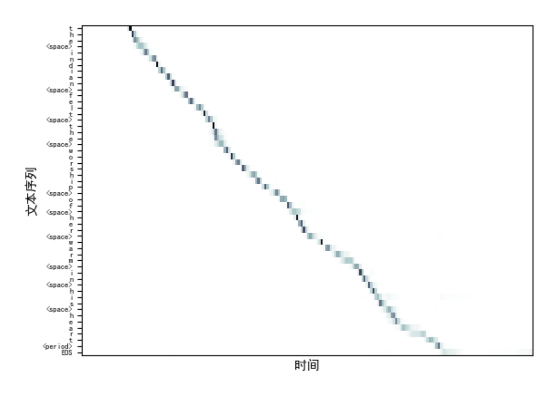
(a)LAS 模型的注意力得分分布图

(a)Attention scores distribution produced by LAS model



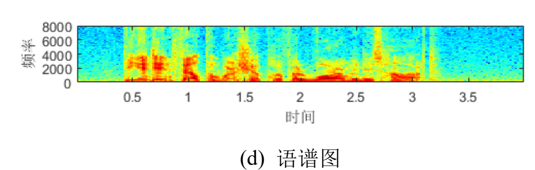
(b)基于卷积位置信息的 LAS 模型的注意力得分分布图

(b)Attention scores distribution produced by LAS model based on convolution location information



(c)新的模型的注意力得分分布图

(c) Attention scores distribution produced by my model



(d) 语谱图

(d)Spectrogram of the signal

图 4　基于同一语音信号(图 4(d))的各模型注意力得分分布对比

Fig.4　Comparison of attention scores for each model based on the same speech signal(Fig.4d))

BLSTM-CTC 模型在纯净和含噪语音测试集上均取得比原始 LAS 模型和基于卷积位置信息的编码器—解码器模型更低的标签错误率，这是因为编码器—解码器模型在当前时刻的解码需要前一时刻的解码信息，当前一时刻的解码出现错误时将会影响当前及未来时刻的解码。寻找比 Scheduled Sampling 更好的训练方法来缓和该种影响，将是未来的研究工作。

参考文献：

[1] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521 (7553): 436.

[2] Hinton G, Deng Li, Yu Dong, *et al.* Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups [J]. IEEE Signal Processing Magazine, 2012, 29 (6): 82-97.

[3] Morgan N, Bourlard H. Continuous speech recognition using multilayer perceptrons with hidden Markov models [C]// Proc of International Conference on Acoustics, Speech, and Signal Processing. [S. l.] : IEEE Press, 1990: 413-416.

[4] Dahl G E, Yu Dong, Deng Li, *et al.* Large vocabulary continuous speech recognition with context-dependent DBN-HMMs [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2011: [S. l.] : IEEE Press, 4688-4691.

[5] Graves A, Fernández S, Gomez F, *et al.* Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]// Proc of the 23rd International Conference on Machine Learning. [S. l.] : ACM Press, 2006: 369-376.

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv, 2014, arXiv: 1409. 0473.

[7] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[8] Cho K, Van Merrienboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2014: 1724-1734.

[9] Chorowski J, Bahdanau D, Cho K, *et al.* End-to-end continuous speech recognition using attention-based recurrent nn: First results [C]// Proc of Workshop on Deep Learning. 2014.

[10] Chan W, Jaitly N, Le Q, *et al.* Listen, attend and spell: a neural network for large vocabulary conversational speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. [S. l.] : IEEE Press, 2016: 4960-4964.

[11] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks [C]// Proc of the 31st International Conference on International Conference on Machine Learning. 2014: II-1764.

[12] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM [C]// Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. [S. l.] : IEEE Press, 2013: 273-278.

[13] Bengio S, Vinyals O, Jaitly N, *et al.* Scheduled sampling for sequence prediction with recurrent neural networks [C]// Advances in Neural



Information Processing Systems. 2015: 1171-1179.

[14] Chorowski J K, Bahdanau D, Serdyuk D, *et al.* Attention-based models for speech recognition [C]// Advances in Neural Information Processing Systems. 2015: 577-585.

[15] Lin M, Chen Q, Yan S. Network in network [J]. arXiv, 2013, arXiv: 1312. 4400.

[16] Graves A, Wayne G, Danihelka I. Neural turing machines [J]. arXiv, 2014, arXiv: 1410. 5401.

[17] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [J]. arXiv, 2015, arXiv: 1508. 04025.

[18] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Advances in neural information processing systems. 2014: 3104-3112.

[19] Surhone L M, Timpledon M T, Marseken S F. Voxforge [M]. 2010.

[20] Varga A, Steeneken H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems [J]. Speech Communication, 1993, 12 (3): 247-251.

[21] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [J]. arXiv, 2015, arXiv: 1502. 03167.

[22] Ba J L, Kiros J R, Hinton G E. Layer normalization [J]. arXiv, 2016, arXiv: 1607. 06450.

[23] Kingma D P, Ba J. Adam: a method for stochastic optimization [J]. arXiv, 2014, arXiv: 1412. 6980.